Detection of Lexical Semantic Changes in Twitter Using Character and Word Embeddings

Yihong Liu
Graduate School of Comprehensive Human Sciences,
University of Tsukuba
s2021716@s.tsukuba.ac.jp

Yohei Seki
Faculty of Library, Information and Media Science,
University of Tsukuba
yohei@slis.tsukuba.ac.jp

1 Introduction

The prosperity of social media platforms on the Internet, such as Twitter, allow users from different countries, even different races and languages, to communicate with each other. Such huge diversity has given rise to the semantic changes of words on the Internet, namely, internet slang words. Numerous internet slang words which changed (New Semantic Word) or newly created (New Blend Word) by the users keep flooding into social media platforms which would quickly become popular and widely been used and affect our daily life than ever. However, semantic meaning Internet slang words could not be collected into the dictionary as well as a corpus in time, which is essential for people or machines to understand them.

Considerable progress has been made in English internet slang words. However, only few works focus on processing Japanese internet slang words [9], and word embeddings methods use only the word as the basic unit and learn embeddings according to the external context of the word, which only contains limited word-level contextual information.

Our proposed method is to combine the character and word embeddings using word2vec and ELMo, respectively. Then we use combination of obtained token representations as the input of the Language Models to detect whether the word in the context is an internet slang word or not.

As there are no public Japanese slang words dataset available, we constructed an internet slang word dataset which contains 40 internet slang words and their meanings as an internet slang word. We separated them into two categories: New Semantic Words and New Blend Words with their characteristics manually. The conducted experiment using this dataset shows that our proposed embedding methods performed better than baseline methods significantly, in processing Japanese internet slang words.

The contributions in this paper are summarized as follows: (1) We proposed a dataset with a novel approach which classifies Japanese internet slang words into two categories, "New Semantic Word", and "New Blend Word"; (2) to obtain token representations for Japanese internet slang words effectively, we also proposed a novel embedding method which combines character and word embeddings utilizing word2vec and ELMo respectively; (3) our experimental results revealed that our proposed encoder which encoded richer syntactic and semantic information about words in-context performed better than baseline methods which used word2vec word embeddings.

This paper is organized as follows. In Section 2, we summarize related works to distinguish Japanese non-standard usages with contextual word embeddings and handling OOV (out-of-vocabulary) words in other Languages by character embeddings. Section 3 describes the Japanese internet slang words dataset we constructed from Twitter. In Section 4, we introduce our proposed model to address the aforementioned problems. In Section 5, the details of the conducted experiments are provided. Finally, the conclusion and our future works are given in Section 6.

2 Related Work

2.1 Internet Slang Words

The emergence of Internet slang words is the result of language evolution. Linguistic variation is a core concept of sociolinguistics [1], a new type of language feature, and a manifestation of the popular culture of social networks. In the information transmitted on social media, some internet slang words have the kind of words with different meanings

from the dictionary, and there are also words that are not listed in the general dictionaries. Therefore, based on this feature, this study divided them into two categories, one is a new semantic word with different meanings from those recorded in dictionaries, and the other is a new blend word that is not recorded in general dictionaries.

Although internet slang words are different from standard words in meaning and usage, they have their own fixed collocations. Therefore, considering these between words in the context, we can extract the words used as internet slang due to the contextual difference. Aoki et al. [9] attempted to distinguish the non-standard usage of common words on social media that are the same as the new semantic words in our research by using contextual word embedding.

2.2 Character-based Word Embeddings

Character-based word representations are now a standard part of neural architectures for natural language processing. Lample et al. [4] illustrated that character representation can be used to handle OOV (out-of-vocabulary) words in a supervised tagging task. Chen et al. [2] proposed multiple-prototype character embeddings to address the issues of character ambiguity and non-compositional words. With the character and word representation, language models will have a more powerful capability of encoding internal contextual information.

3 Internet Slang Corpus

In this section, we describe the construction of our dataset in Section 3.1, and our prepossessing on the dataset by separating extracted Japanese internet slang words into two proposed types are described in Section 3.2 and 3.3.

3.1 Corpus Construction

Dataset in our experiment is constructed using crawled tweets in Japanese via Twitter API. First, we defined 40 internet slang words, while 20 words belong to New Semantic Word and 20 words belong to New Blend Word. The details of New Semantic Word and New Blend Word are given in Section 3.2 and 3.3, respectively. Then we extracted 30 sentences that contain the internet slang words, while another 30 sentences also contain the same words but used as not internet slang for a comparison. Thirdly, the collected Japanese texts are segmented in two ways, one is to separate the characters based on the character-level, and the other is based on the word unit to segment words by Mecab[1)]. Finally, we explain the annotation tags for tokens. For word-level, *sem* stand for New Semantic Word, *bln* means New Blend Word, and *O(Others)* stand for common words. While in character-level, we also add *BIO(B-Begin I-Inside O-Others)* tagging style to represent the position information of character-based tokens. Examples are given in Table 1.

Table 1 Annotations of Japanese Internet Slang Words

New Semantic Word
-Character-level
初/ O 鯖/B-sem の/O 初/O 心/O 者/O に/O
迷/O 惑/O か/O け/O る/O な/O ！/O
-Word-level
初/ O 鯖/sem の/O 初心者/O に/O
迷惑/O かける/O な/O ！/O
New Blend Word
-Character-level
そ/B-bln マ/I-bln ？/O 行/O け/O る/O 時/O
言/O っ/O て/O バ/O イ/O ト/O 無/O け/O れ/O
ば/O ワ/O イ/O も/O 行/O く/O わ/O
-Word-level
そマ/bln ？/O 行ける/O 時/O 言っ/O て/O
バイト/O 無けれ/O ば/O ワイ/O も/O 行く/O わ/O

3.2 New Semantic Word

New Semantic Words are often the vocabulary that is originally recorded in the dictionary and used daily. It, however, has new meanings which used widely because of its similarity in pronunciation to other terms, or some iconic popular events, etc. Although this type of vocabulary can be segmented directly, the meaning of the text will be very different because of the change in meaning and even part of speech. Examples are shown in the Table 2,

Table 2 Examples of New Semantic Word

New Semantic Word
Internet Slang Word: Meaning/ English Translation
-**草**: 面白い/interesting
-**丸い**: 無難/safe
-**炎上**: 批判のコメントが殺到する状態/internet troll

1) https://github.com/neologd/mecab-ipadic-neologd

Table 3 Examples of New Blend Word

New Blend Word
Internet Slang Word: Meaning/ English Translation
-**わかりみ**: 分かること/understanding
-**禿同**: 激しく同意/strongly agree
-**ふぁぼ**: お気に入り/favourite

3.3 New Blend Word

New Blend Words often borrow foreign words, dialects, numeric elements, and icons. They often combine definitions, homonyms, abbreviations, repetitions, and other word formation methods as well as unconventional grammars [3]. Internet language has achieved the effect of "novelty" through its unconventional nature and non-standard usage in its defining characteristic. An example is given in the Table 3.

4 Proposed Encoder Model

We propose an encoder model which combined the character and word embeddings for each token as the token representation. We take them as the input for two layers of biLSTM network, which will be accumulated to the sum according to their respective weights by referring to [5].

Considering the relationships between the word and its characters, the discriminative characters are crucial for distinguishing the slight semantic differences [7]. In terms of the joint embedding model, word embedding can store textual information among words. Adding character representation also provide the semantic information between characters.

The embedding obtained by word2vec is fixed, while the ELMo [6] can change it flexibly according to the context. Therefore, using embeddings from word2vec which obtain static standard contextual information and accumulating them into ELMo representation can make the semantic distinction based on the context from the dataset.

Following Rei et al. [8], we also uses CRF as the output layer to predict token's tag. In this architecture, the last hidden layer is used to predict confidence scores for the word with each of the possible labels.

4.1 Character and Word Embedding

The general structure of our model is given in Fig. 1.

The parameter w_j is the word embedding of token j, N_j is the number of characters in this token, c_j^k is the embedding of the k-th character. x_j and y_j^k are joint embeddings for the Word-level unit annotation and Character-level unit annotation, which equations are given as (1) and (2), respectively:

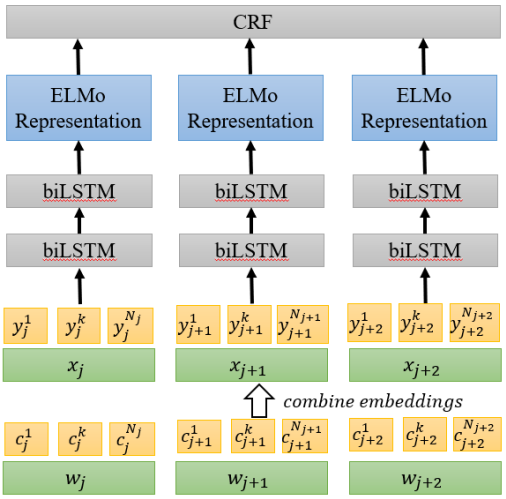

Figure 1 Structure of Our Proposed Model

$$x_j = w_j \oplus \frac{1}{N_j} \sum_{k=1}^{N_j} c_j^k \tag{1}$$

$$y_j^k = c_j^k \oplus \frac{w_j}{N_j} \tag{2}$$

4.2 ELMo Representation

We also trained the ELMo model introduced by Peter et al. [6] with the same Japanese Wikipedia Dataset, whereby all of the layers are combined with a weighted average pooling operation.

$$\mathbf{ELMo}_k = \gamma \sum_{j=0}^{L} s_j \mathbf{h}_{k,j} \tag{3}$$

where $\mathbf{h}_{k,j}$ is the hidden state outputs for each biLSTM layer. The parameters s are softmax-normalized weights and scalar parameter γ allows the task model to scale the entire ELMo vector. γ means practical importance to aid the optimization process.

5 Experiment

5.1 Dataset and Setting

We selected the Japanese Wikipedia Dataset[2] for character and word embedding learning where the number of words and characters wre about 312 thousands and 10 thousands respectively. We set vector dimension as 200 and context window size as 10 for character based and 5 for word based, and also set other setting as same as in the Japanese Wikipedia Entity Vector[3].

5.2 Baseline Methods

To verify the superiority of the joint embedding model of words and characters and the ability to distinguish semantics with ELMo embedding, we set two baseline models as follows.

(1) Baseline1 (w): word-only embeddings with LSTM networks (2) Baseline2 (w+c): character and word embeddings with LSTM networks

Meanwhile, we conduct two-sided t-Test for the average value of the precision, recall, and F1 score for 10 words in two annotation levels so as to investigate the statistical significance between our model and two baseline models.

5.3 Result

Experimental results on New Semantic Word and New Blend Word are shown in Table 4 and in Table 5.

Our results proved that the joint embedding model has a significant improvement in the results to extract internet slang words compared to the baselines using word embedding only, or using LSTM instead of ELMo. Of course, due to the complexity of the embedding model, in the experiment, the model using joint embedding converges to the same minimum loss, which requires more iterations. In particular, considering that the dimension of ELMo itself is much larger than the basic bidirectional LSTM (200 dimensions vs. 1024 dimensions), the iteration of the ELMo model requires more layers of iteration.

Regardless of whether it is a new semantic word or a new blend word, our model worked effectively, especially for the word level annotation corpus. This may be due to the fact that the semantic information between characters is too rich and complicated in the text annotated in character units, especially for new semantic words. It is obvious that the results of the new blend words are better under the character unit level annotation.

In addition, because most of the new blend words are innovative words, there is less semantic change, and most of them are accompanied by fixed collocations, or have a fixed place in the sentence. Combining all the experiments and methods, the extraction methods of our model for such words performs better.

2) https://dumps.wikimedia.org/jawiki/latest/, accessed on Oct 2020
3) http://www.cl.ecei.tohoku.ac.jp/ m-suzuki/jawiki_vector/

Table 4 Results of Detecting New Semantic Words

Model	Level	Precision	Recall	F1-score
Baseline1 (w)	word	0.045	0.063	0.085
Baseline2 (w+c)	word	0.245	0.844	0.380
Our model	word	$0.618^{**,-}$	$0.872^{**,-}$	$0.723^{**,-}$
Baseline1 (w)	char	0.024	0.625	0.047
Baseline2 (w+c)	char	0.112	0.877	0.198
Our model	char	$0.170^{*,-}$	$0.714^{**,*}$	$0.275^{*,-}$

Table 5 Results of Detecting New Blend Words

Model	Level	Precision	Recall	F1-score
Baseline1 (w)	word	0.439	0.678	0.532
Baseline2 (w+c)	word	0.576	0.795	0.667
Our model	word	$0.878^{**,**}$	$0.884^{-,-}$	$0.881^{**,*}$
Baseline1 (w)	char	0.385	0.343	0.363
Baseline2 (w+c)	char	0.444	0.369	0.403
Our model	char	$0.576^{-,**}$	$0.312^{*,*}$	$0.405^{-,**}$

* Compared to the baselines, our model cases improved at the significance level of 5%.
** Compared to the baselines, our model improved at the significance level of 1%.
- Compared to the baselines, our model improved but with no significance. The left symbol means the significant difference for Baseline1, while the right symbol for Baseline2.

6 Conclusion

We have shown that the combination of character and word embeddings through the deep bidirectional Language Models can learn the rich information from the context, which not only contain the association between characters and words, but also utilize contextual semantic information of characters effectively. In future work, we will extract the meaning of the internet slang words by joint embedding models.

Acknowledgement

This work was partially supported by a JSPS Grant-in-Aid for Scientific Research (B) (#19H04420).

References

[1]Jack K Chambers. *Sociolinguistic Theory*. Wiley-Blackwell, 2008.

[2]Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huanbo Luan. Joint learning of character and word embeddings. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, p. 1236–1242, 2015.

[3]Fazal Masud Kundi, Shakeel Ahmad, Aurangzeb Khan, and Muhammad Zubair Asghar. Detection and scoring of internet slangs for sentiment analysis using sentiwordnet. *Life Science Journal*, Vol. 11, No. 9, pp. 66–72, 2014.

[4]Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270, San Diego, California, June 2016.

[5]Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1756–1765, Vancouver, Canada, July 2017.

[6]Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, June 2018.

[7]Xue Qiao, Chen Peng, Zhen Liu, and Yanfeng Hu. Word-character attention model for chinese text classification. *International Journal of Machine Learning and Cybernetics*, Vol. 10, No. 12, pp. 3521–3537, 2019.

[8]Marek Rei, Gamal Crichton, and Sampo Pyysalo. Attending to characters in neural sequence labeling models. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 309–318, Osaka, Japan, December 2016.

[9]青木竜哉, 笹野遼平, 高村大也, 奥村学. ソーシャルメディアにおける単語の一般的ではない用法の検出. 自然言語処理, Vol. 26, No. 2, pp. 381–406, 2019.